Evolution of Interconnects for Supercomputing

Moray McLaren
Quadrics Ltd.

EtherNet and EtherNot!

Will standard interconnects solve all our problems?

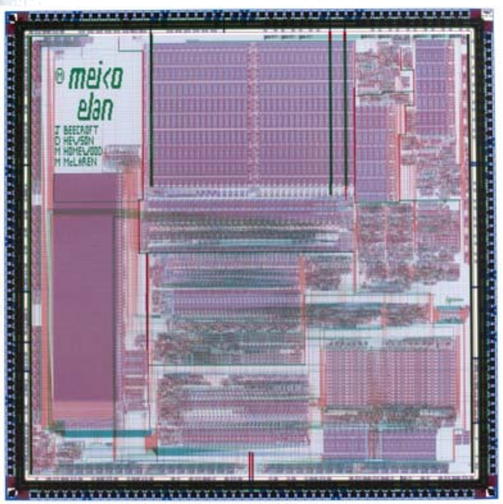
- Whatever the volume interconnect of the future is, it will be called Ethernet.
- Incorporate ideas from specialised low latency interconnects into Ethernet?
  - RDMA is a start
  - Common DDI with high performance NICs?
  - Price advantage not so clear for equivalent BW.
- Successful EtherNot technologies need clear performance advantages that deliver in applications.

What’s special about Supercomputing?

- You push the extremes of scale
  - Seamless switch scaling
  - Global operations
  - Fault tolerance
- You value your compute cycles
  - Compute communications ratio
  - Ultra low latency
  - Overhead

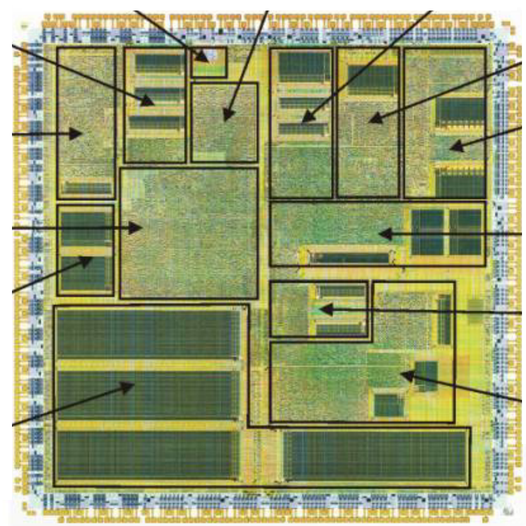
Historical scaling…

Elan – 1990



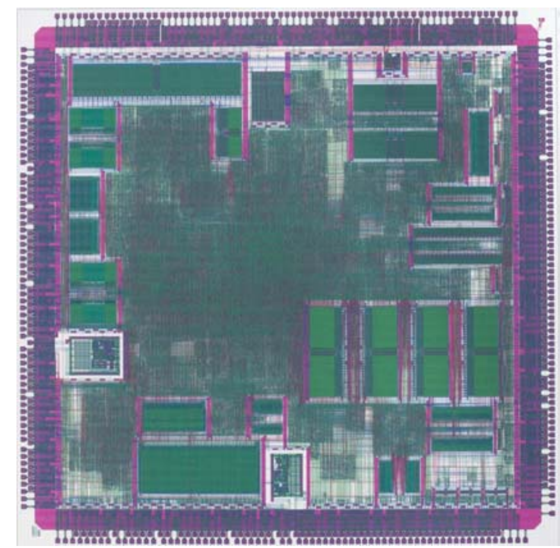
Put - 9µs
MPI - 78µs
44Mbytes/s

Elan 3 – 1998



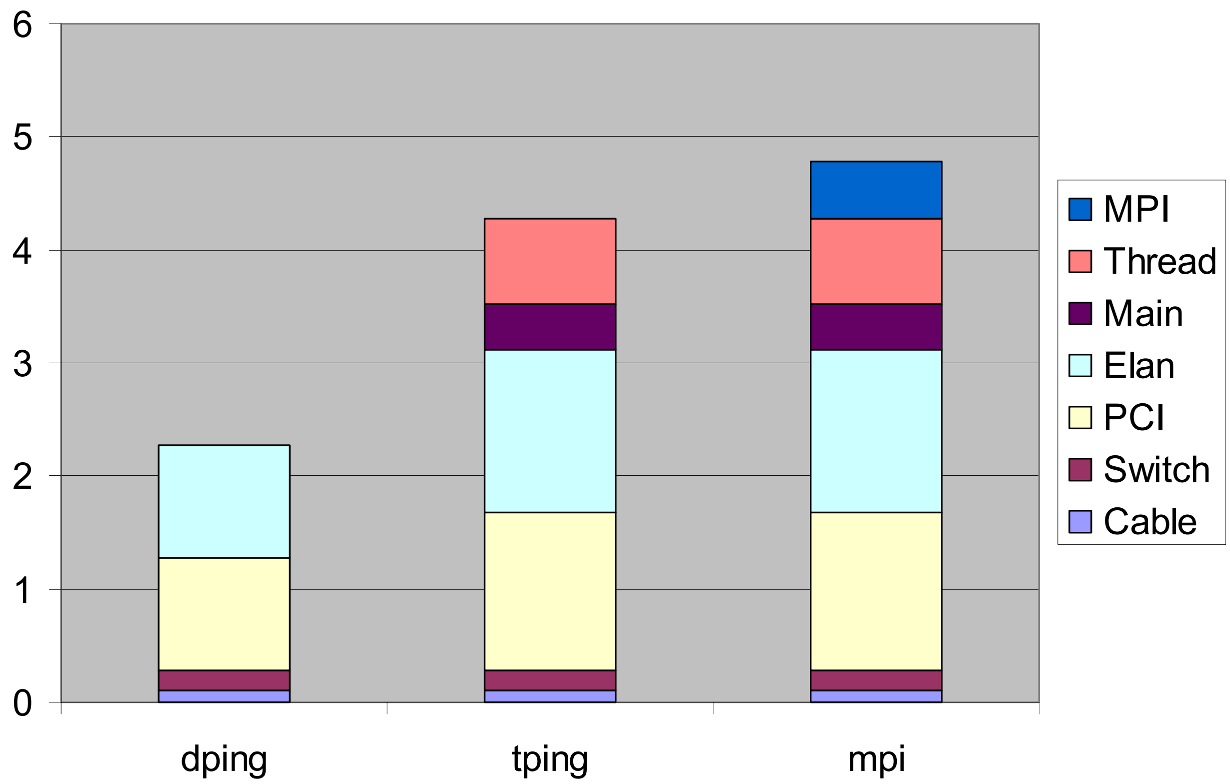
Put - 2µs
MPI - 5µs
320Mbytes/s

Elan 4 – 2003

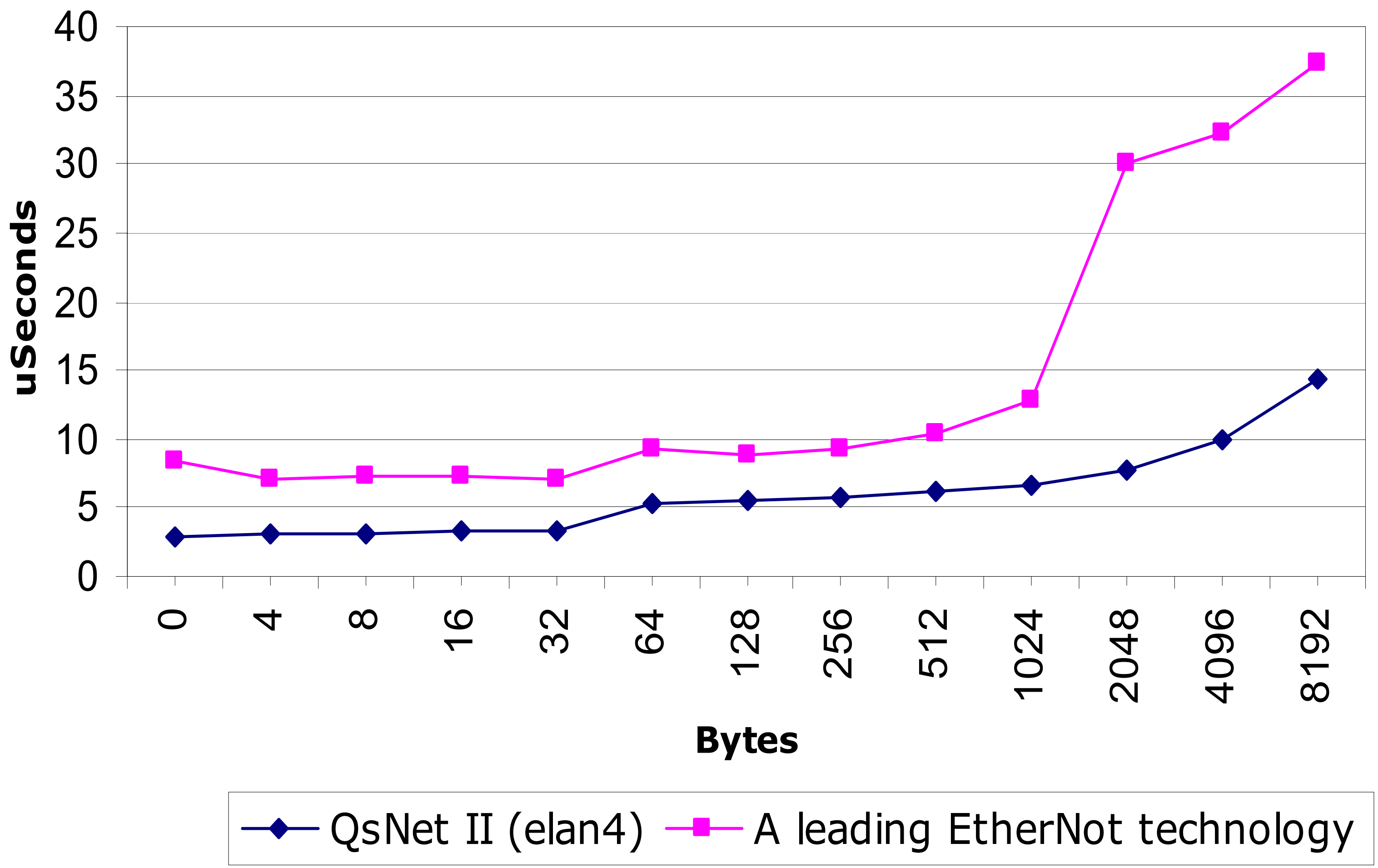


Put - 2µs
MPI - 3µs
900Mbytes/s

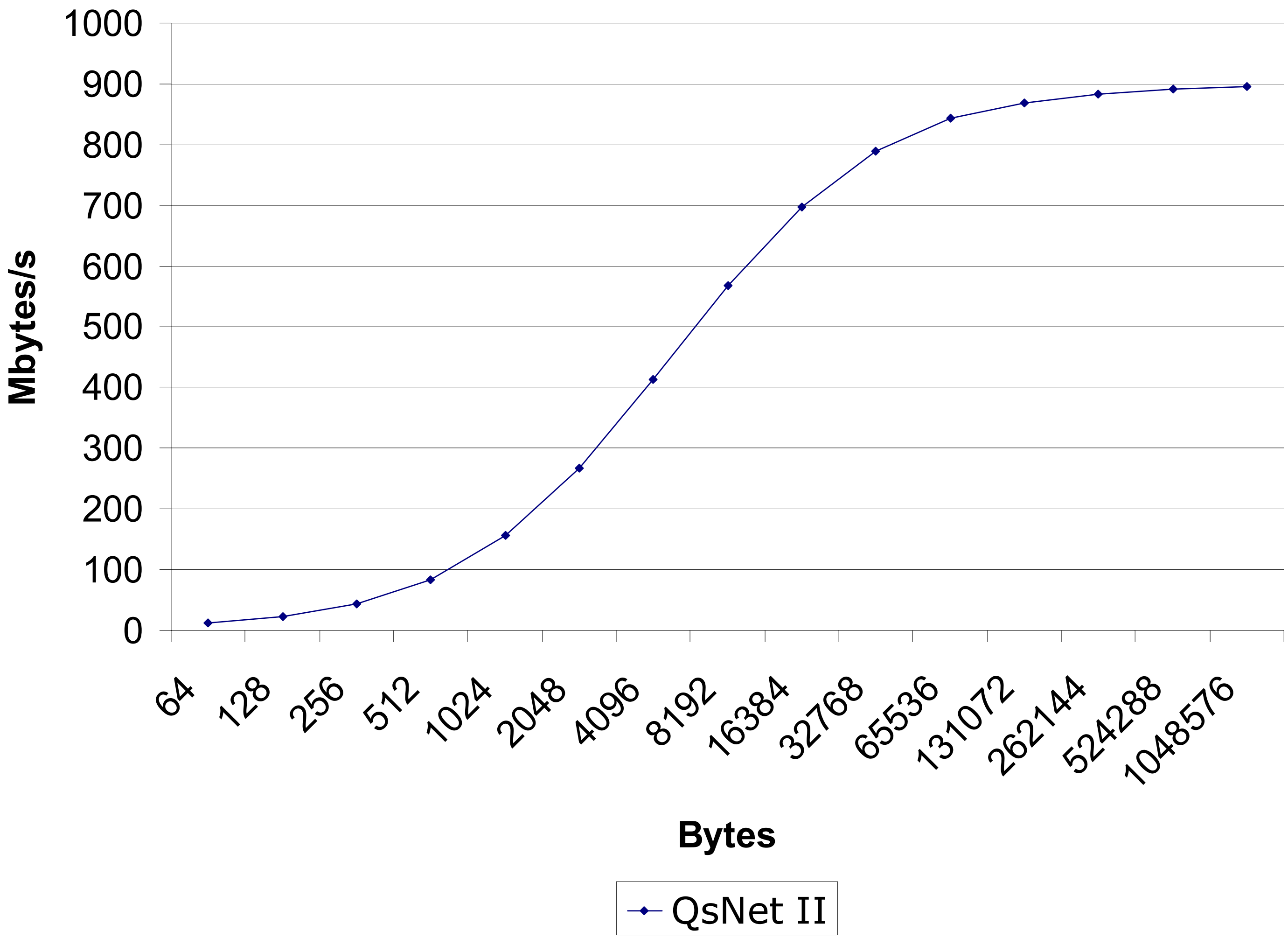
Elan-3 MPI Latency Breakdown



MPI short message latency



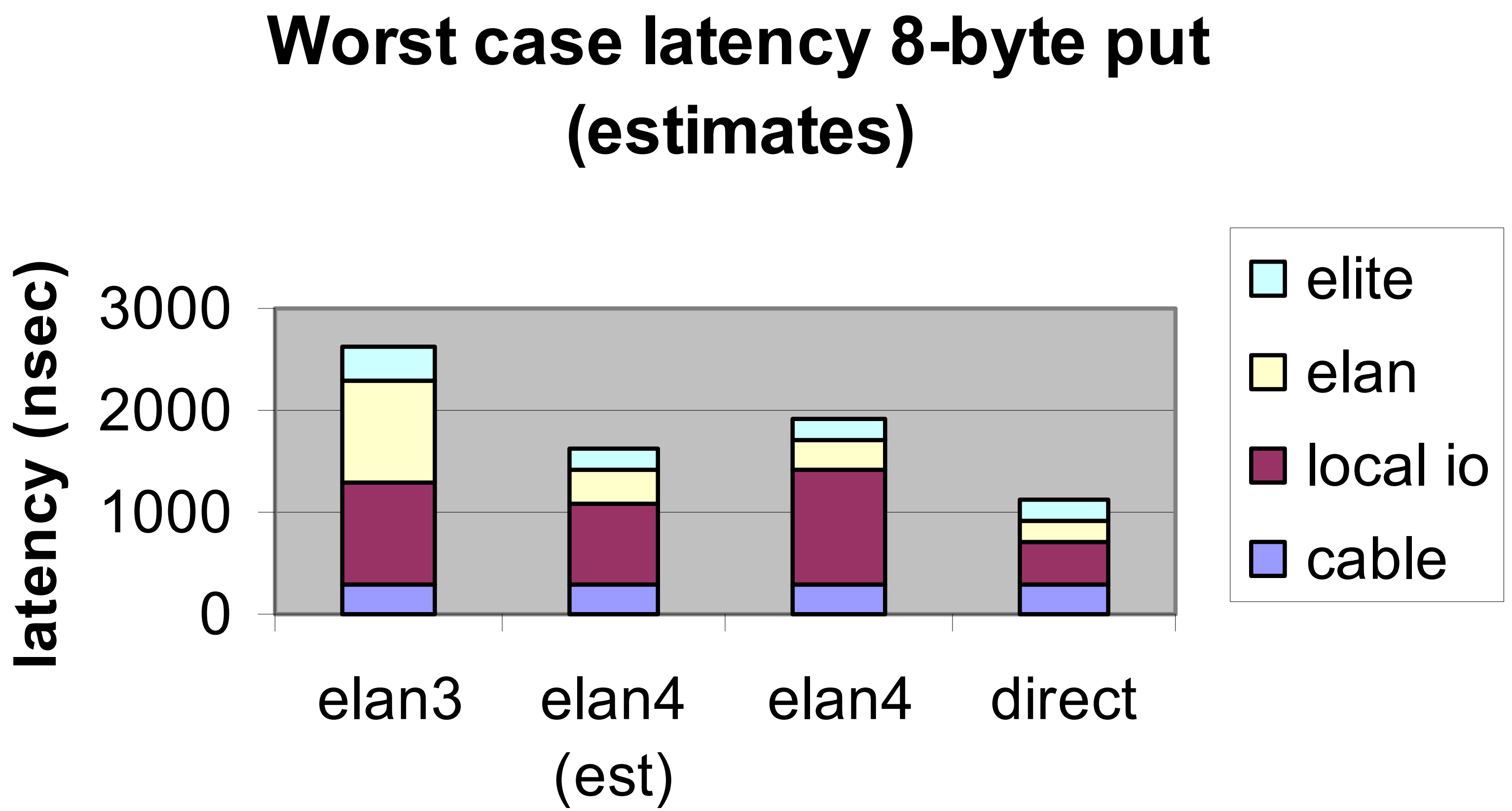
MPI Bandwidth – Elan 4



Bye bye MPI?

- Problems with MPI
  - High overhead for very short messages
  - Tag matching overhead
  - MPI ordering rules imply single point of ordering for each node.
- remote read, remote write API
  - draw on libElan, ShMem
  - Support for many outstanding transactions
  - Target for compiler writers, and library developers

Low level hardware latency



Getting closer to the CPU…

- Where?
  - Hypertransport, proprietary IO port
- What's the win?
  - Avoid bus bridge latency
  - Lower cache refill overhead – maybe?
  - Simpler interface
  - Smaller transfers for peak performance.
- Issues – primarily commercial
  - Where's the connector?
  - Fragmentation of silicon volumes

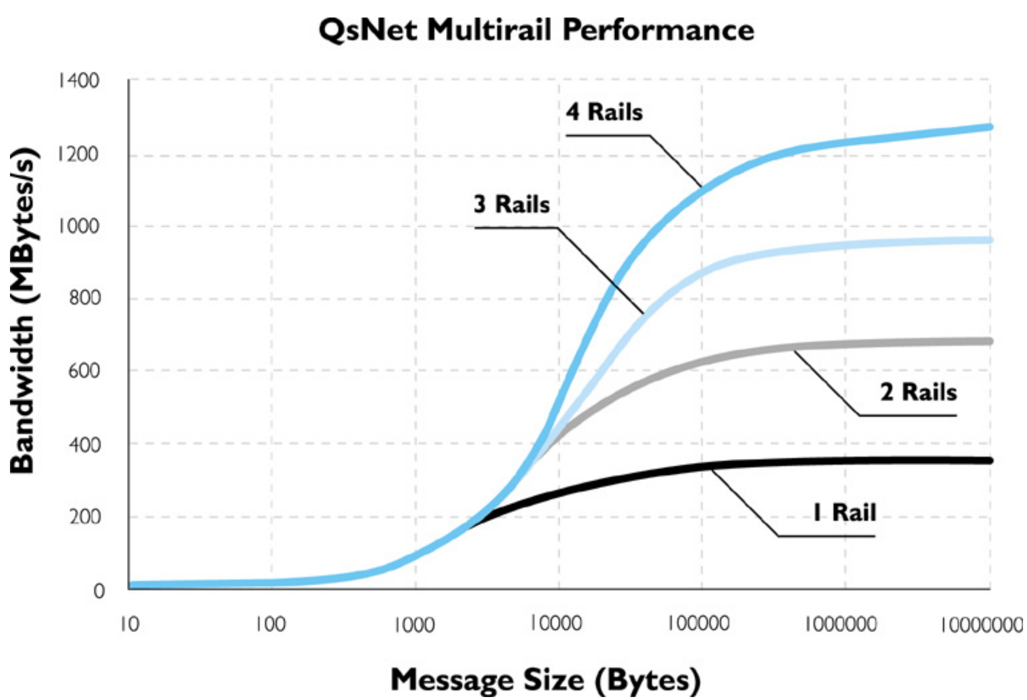
The way forward on latency

- Basic hardware latency
  - Many factors reaching practical limits.
  - Closer integration to CPU removes some delays
  - Pipelining to support multiple outstanding short messages
- Real application latency
  - MPI well understood
  - Lower level API need for compilers etc.
  - What's the API for kernel messaging?
  - Reliably, ordered, datagram.
  - Several alternates, Portals, Via constructs..

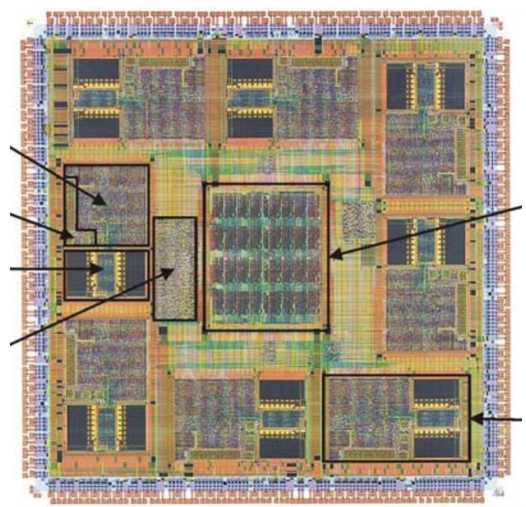
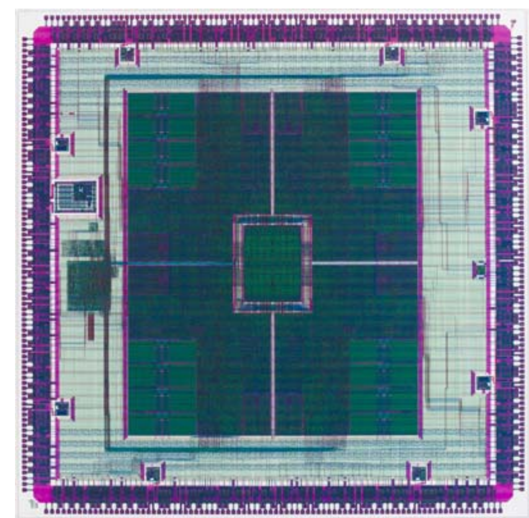
Bandwidth going forward

- Limited by where you can connect to
  - Double and Quad clock PCI-X
  - PCI-Express
  - Direct connections.
- Large scale multi rail systems with large SMPs
  - NUMA challenges
  - Separate rails or one big switch?

QsNet (Elan 3) Multirail Performance

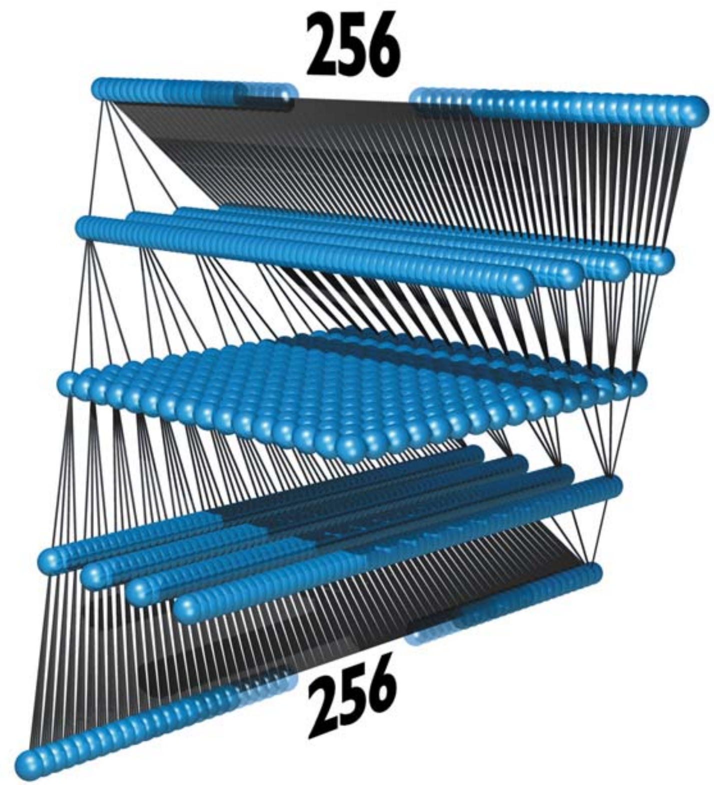


Switch scaling

Elite – 1990	Elite 3 – 1998	Elite 4 – 2003
		
70Mhz	**400Mhz**	**1.3GHz**
44Mbytes/s	**325Mbytes/s**	**900Mbytes/s**
256 nodes	**~2k nodes**	**~4K**

Topology

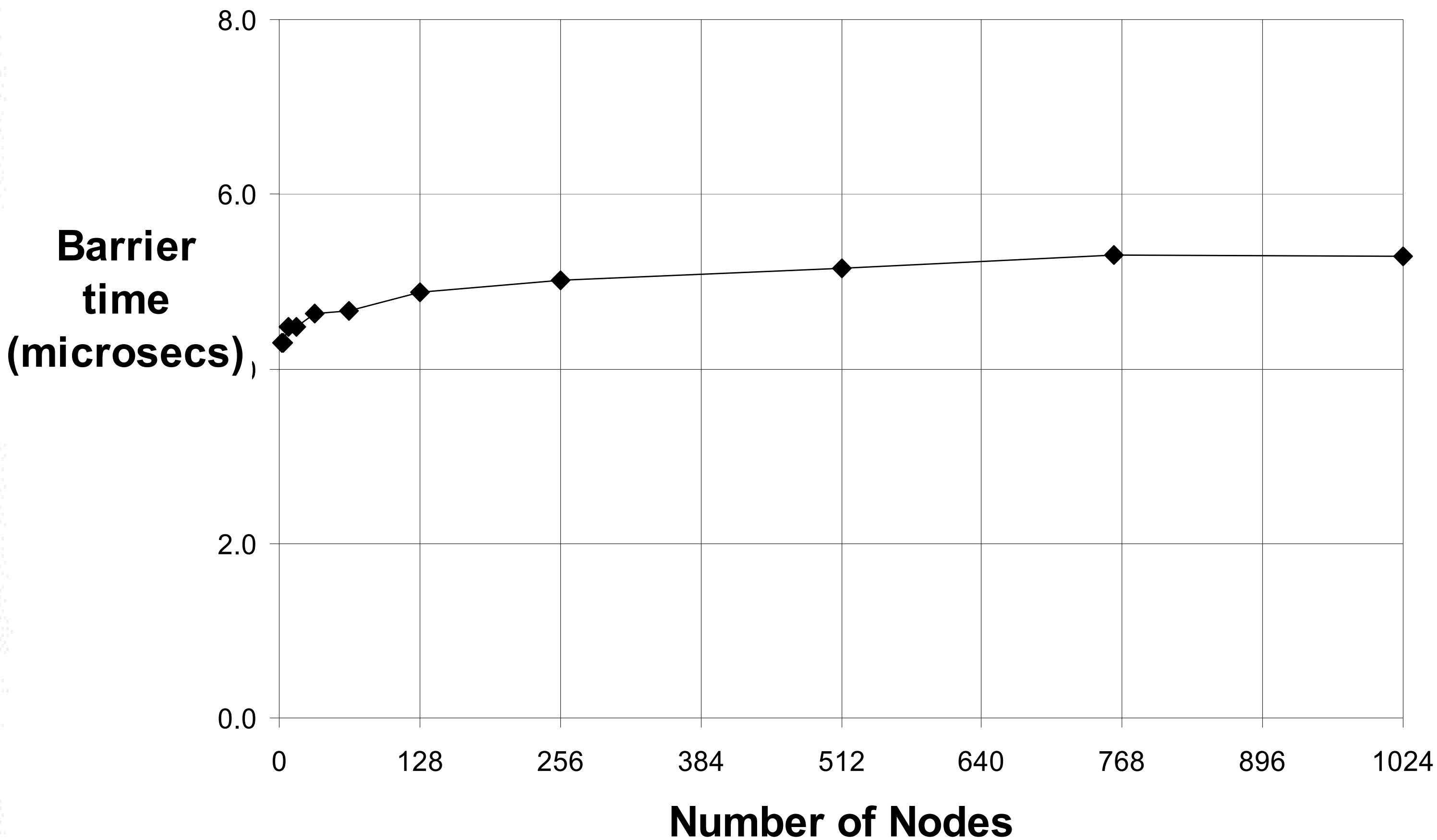
- Fat trees have been very successful
  - Good structure for fault tolerance
  - Fairly uniform connectivity.
  - Good for global operations
  - Quite challenging to for systems integrations
- Packaging issues will dictate topology



Global operations

- Why do they matter?
  - Improve application scaling on very large systems
  - Highly scalable single system image functions, (e.g. cluster membership)
- What works now
  - Range selected broadcast, barrier in the network
  - Integer collectives handled by IO processor in NIC
- Future
  - Floating point collectives (probably more appropriate in the NIC than in the switch)
  - Alternative broadcast constructs.

Barrier Scaling (QsNet)



Data Courtesy of Lawrence Livermore National Lab

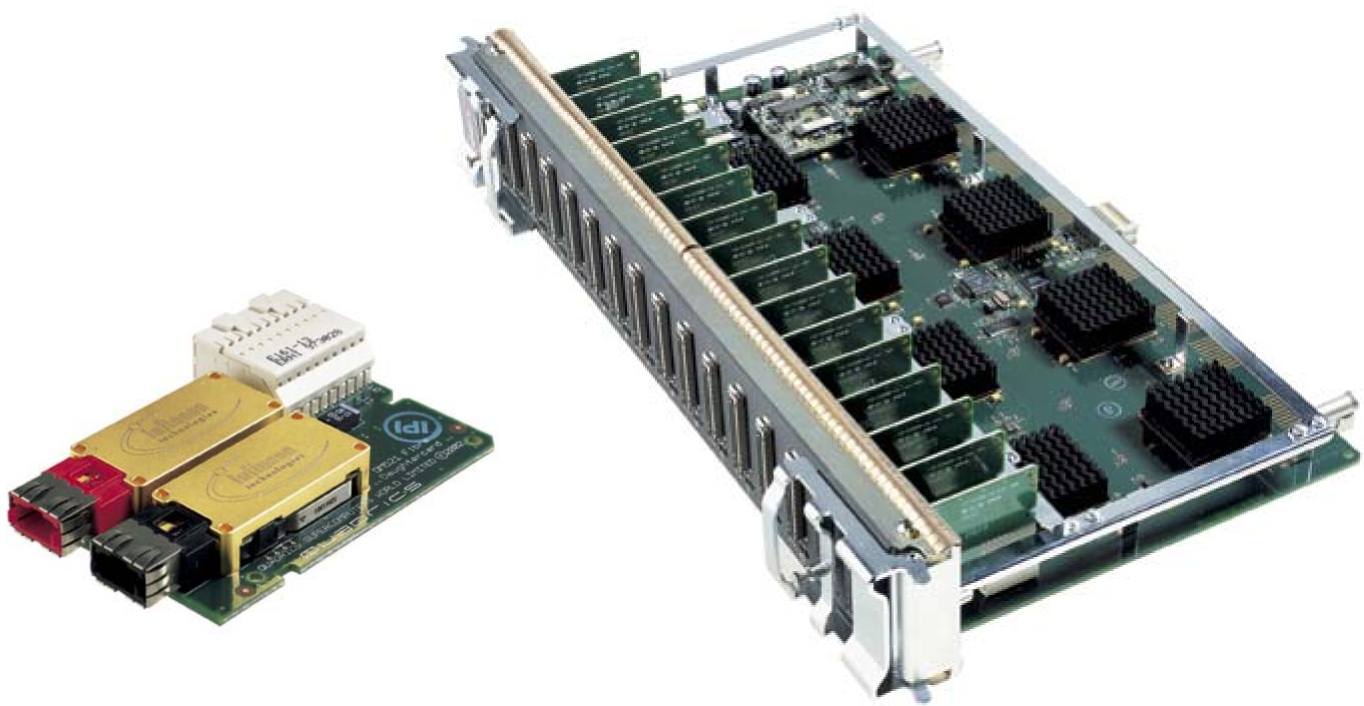
Fault tolerance and availability

- Typical current features for RAS
  - Use a topology with lots of alternate routes
  - Dual redundant PSU
  - Tolerant to single fan failure
  - Dual redundant control cards

Development for fault tolerance & reliability

- Pairwise broadcast for hardware mirroring.
- Rail failover in multirail systems
- Best way to improve interconnect reliability is minimise connectors

Physical packaging considerations



QsNet^{II} Physical Link

- 1.333Ghz design speed
  - 4b5b coding for DC balance
  - ~900 Mbytes/s after protocol
- Copper
  - 10 bit lvds – total 40 wires
  - 10-12m range
- Optics
  - 12 bit parallel optical fiber
  - 100m

Future link technologies

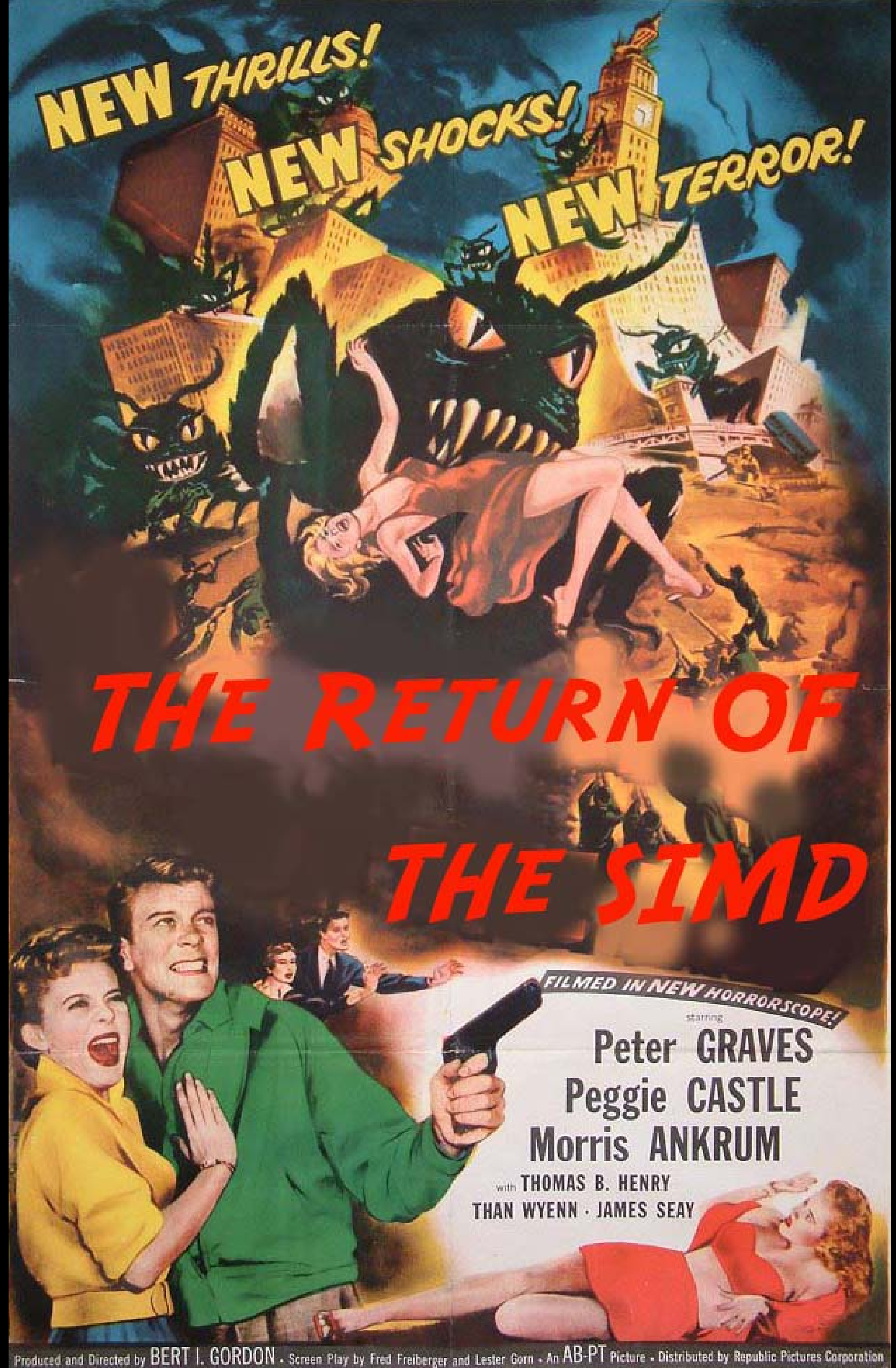
- Still copper on the backplane for cost and reliability
  - Careful design gets to up to 5Gbit/s per wire for moderate runs. More with clever equalisation.
  - Max length decreases as speed increase
  - Improved packing technology to reduce connection lengths, pack more into the copper zone.
- Rack to rack all fibre
  - Future generations of parallel fibre
  - 12 x 5 Gbits/s ~= 6Gbytes/s

Optical switching ?

- Optical technology has been driven by telecomms requirements
  - Long haul not short haul
  - Circuit switched not packet switched
  - They’re not buying anything!
- Combining logic, switching and buffering – easy for silicon
  - Silicon switches – a distributed arbiter which delivers data as a side effect.

Optical switching?

- Optical crossbars or WDM based
- Advantages
  - Large crossbars possible
  - Very low latency for established connections.
  - Integration with optical fibre
- But…
  - New component technologies
  - Optical switching generally have a separate control plane
  - Difficult to build self routing packet switches
  - Implies switch architecture with centralised control
- What sort of machines can we make with these?


NEW THRILLS!
NEW SHOCKS!
NEW TERROR!
THE RETURN OF
THE SIMD
FILMED IN NEW HORRORSCOPE!
starring
Peter GRAVES
Peggie CASTLE
Morris ANKRUM
with THOMAS B. HENRY
THAN WYENN · JAMES SEAY
Produced and Directed by BERT I. GORDON • Screen Play by Fred Freiberger and Lester Gorn • An AB-PT Picture • Distributed by Republic Pictures Corporation